

A Projection Based Conditional Dependence Measure with Applications to High-dimensional Undirected Graphical Models

Jianqing Fan^[1], Yang Feng^[2] and Lucy Xia^[3]

^[1] *Department of Operations Research & Financial Engineering, Princeton University, Princeton, New Jersey 08544, U.S.A.*

^[2] *Department of Statistics, Columbia University, New York, NY 10027, U.S.A.*

^[3] *Department of Statistics, Stanford University, Stanford, CA 94305, U.S.A.*

Summary. Measuring conditional dependence is an important topic in statistics with broad applications including graphical models. Under a factor model setting, a new conditional dependence measure based on projection is proposed. The corresponding conditional independence test is developed with the asymptotic null distribution unveiled where the number of factors could be high-dimensional. It is also shown that the new test has control over the asymptotic significance level and can be calculated efficiently. A generic method for building dependency graphs without Gaussian assumption using the new test is elaborated. Numerical results and real data analysis show the superiority of the new method.

Keywords: conditional dependency; distance covariance; factor model; graphical model; projection

1. Introduction

Undirected graphical model is an important tool to capture dependence among random variables and has drawn tremendous attention in various fields including signal processing, bioinformatics and network modeling (Wainwright and Jordan, 2008). Let $\mathbf{z} = (z^{(1)}, \dots, z^{(d)})$ be a d -dimensional random vector. We denote the undirected graph corresponding to \mathbf{z} by (V, E) , where vertices V correspond to components of \mathbf{z} and edges $E = \{e_{ij}, 1 \leq i, j \leq d\}$ indicate whether node $z^{(i)}$ and $z^{(j)}$ are conditionally independent given the remaining nodes. In particular, the edge e_{ij} is absent if and only if $z^{(i)} \perp\!\!\!\perp z^{(j)} | \mathbf{z} \setminus \{z^{(i)}, z^{(j)}\}$. When \mathbf{z} follows multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the precision matrix $\boldsymbol{\Omega} = (w_{ij})_{d \times d} = \boldsymbol{\Sigma}^{-1}$ captures exactly this relationship; that is, $w_{ij} = 0$ if and only if e_{ij} is absent (Lauritzen, 1996; Edwards, 2000). Therefore, under the Gaussian assumption, this problem reduces to the estimation of precision matrix, where a rich literature on model selection and parameter estimation can be found in both low-dimensional and high-dimensional settings, including Dempster (1972), Drton and Perlman (2004), Meinshausen and Bühlmann (2006), Friedman *et al.* (2008), Fan *et al.* (2009), and Cai *et al.* (2011). While Gaussian graphical model (GGM) can be useful, the stringent requirement on normality is not always satisfied in real application where the observed data usually have fat tails or are skewed (Xue and Zou, 2012).

To relax the Gaussian assumption, Liu *et al.* (2009) proposed the nonparanormal model, where they find transformations that marginally gaussianize the data and then work under the Gaussian graphical model framework to estimate the network structure. Under the nonparanormal model, Xue and Zou (2012) proposed rank-based estimators to approximate the precision matrix. The nonparanormal model, although flexible, still assume the transformed data follows a multivariate Gaussian distribution, which can also be restrictive at times. Instead of using these nonparametric methods to find transformations and

work under GGM, we would like to propose a more natural way of constructing graphs. That is, we work directly on the conditional dependence structure by introducing a measure for conditional dependence between node i and j given the remaining nodes. Then, we can introduce a hypothesis testing procedure to decide whether the edge e_{ij} is present or not.

It is worth noting that, based on hypothesis testing, we could indeed build a general conditional dependency graph, where presence of an edge e_{ij} between nodes $z^{(i)}, z^{(j)}$ represents that the two nodes are dependent conditional on some factors \mathbf{f} . Graphical model is one type of such graphs where \mathbf{f} is chosen to be $\mathbf{z} \setminus \{z^{(i)}, z^{(j)}\}$; in such cases, we call \mathbf{f} internal factors. More generally, \mathbf{f} could contain covariates we observe or latent factors that we do not observe, which are not necessarily part of \mathbf{z} , and we call them external factors. As an example, in Fama-French three-factor model, the return of each stock can be considered as one node and \mathbf{f} are the chosen three-factors. This example will be further elaborated in Section 5. Another interesting application is discussed in Stock and Watson (2002), where external factors are aggregated macroeconomic variables, and the nodes are disaggregated macroeconomic variables.

Back to hypothesis testing, in economics, there has been abundant literature on different conditional independence tests. Linton and Gozalo (1997) proposed two nonparametric tests of conditional independence based on a generalization of the empirical distribution function; however, a complicated bootstrap procedure is needed to calculate critical values of the test, which leads to limited practical value. Su and White (2007, 2008, 2014) proposed conditional independence tests based on Hellinger distance, conditional characteristic function and empirical likelihood, respectively. However, all those tests either have tuning parameters or are computationally expensive.

As an attempt to propose a remedy for the above issues, we consider the following setup. In particular, suppose $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{f}_i, \epsilon_{i,x}, \epsilon_{i,y}), i = 1, \dots, n\}$ are i.i.d. realizations of $(\mathbf{x}, \mathbf{y}, \mathbf{f}, \epsilon_x, \epsilon_y)$, which are generated from the following model:

$$\mathbf{x} = \mathbf{G}_x(\mathbf{f}) + \epsilon_x, \quad \mathbf{y} = \mathbf{G}_y(\mathbf{f}) + \epsilon_y, \quad (1)$$

where \mathbf{f} is the K -dimensional common factors, \mathbf{G}_x and \mathbf{G}_y are general mappings from \mathbb{R}^K to \mathbb{R}^p and \mathbb{R}^q , respectively. Note that we only observe $\{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{f}_i), i = 1, \dots, n\}$. Here, we assume independence between (ϵ_x, ϵ_y) and \mathbf{f} . In addition, the dimensions p and q are assumed to be fixed while the number of factors K could diverge to infinity.

Our goal is to test whether \mathbf{x} and \mathbf{y} are independent given \mathbf{f} , i.e.,

$$H_0 : \mathbf{x} \perp\!\!\!\perp \mathbf{y} | \mathbf{f}. \quad (2)$$

Under model (1), the testing problem is equivalent to test whether ϵ_x and ϵ_y are independent, i.e.,

$$H_0 : \epsilon_x \perp\!\!\!\perp \epsilon_y. \quad (3)$$

Note that in the case of building graphical models without Gaussian assumption, we could use (1) by setting \mathbf{x} and \mathbf{y} as the random vectors associated with a pair of nodes in the graph and \mathbf{f} represents the rest of the nodes. Then, the method to be developed could be used to construct a high-dimensional undirected graphs by conducting the test in (2) for each pair of the nodes. This gives a graphical summary of the conditional dependence structure.

Since $\{(\epsilon_{i,x}, \epsilon_{i,y}), i = 1, \dots, n\}$ are hidden, a natural idea is to estimate them by the residuals after a projection of \mathbf{x} and \mathbf{y} onto \mathbf{f} . Asymptotically, a fully nonparametric projection on \mathbf{f} (e.g., local polynomial

regression) would consistently recover the random errors when K is fixed along with certain smoothness assumptions on \mathbf{G}_x and \mathbf{G}_y . However, it becomes challenging when K diverges due to the curse of dimensionality if no structural assumptions are made on \mathbf{G}_x and \mathbf{G}_y . As a result, we will study the case where \mathbf{G}_x and \mathbf{G}_y are linear functions (factor models) in Section 2.2 and the case where \mathbf{G}_x and \mathbf{G}_y are additive functions in Section 2.5 when K diverges.

To complete our proposal, we need to find a suitable measure of dependence between random variables/vectors. In this regards, many different measures of dependence have been proposed. Some of them rely heavily on Gaussian assumptions, such as Pearson correlation, which measures linear dependence and the uncorrelatedness is equivalent to independence only when the joint distribution is Gaussian; or Wilks Lambda (Wilks, 1935), where normality is adopted to calculate the likelihood ratio. To deal with non-linear dependence and non-Gaussian distribution, statisticians have proposed rank-based correlation measures, including Spearman's ρ and Kendall's τ , which are more robust than Pearson correlation against deviations from normality. However, these correlation measures are usually only effective for monotone types of dependence. In addition, under the null hypothesis that two variables are independent, no general statistical distribution of the coefficients associated with these measures has been derived. Other related works include Hoeffding (1948), Blomqvist (1950), Blum *et al.* (1961), and some methods described in Hollander *et al.* (2013) and Anderson (1962). Taking these into consideration, distance covariance (Székely *et al.*, 2007) was introduced to address all these deficiencies. The major benefits of distance covariance are: first, zero distance covariance implies independence, and hence it is a true dependence measure. Second, distance covariance can measure the dependence between any two vectors which potentially are of different dimensions. Due to these advantages, we will focus on distance covariance in this paper as our measure of dependence.

The main contribution of this paper is two-fold. First, under the factor model assumption, we propose a computationally efficient conditional independence test. Both the response vectors and the common factors can be of different dimensions and the number of the factors could grow to infinity with sample size. Second, we apply this test to build conditional dependency graph and covariates-adjusted dependency graph, as generalizations of the Gaussian graphical model.

The rest of this paper is organized as follows. In Section 2, we present our new procedure for testing conditional independence via projected distance covariance (P-DCov) and describe how to construct conditional dependency graphs based on the proposed test. Section 3 gives theoretical properties including the asymptotic distribution of the test statistic under the null hypothesis as well as the type I error guarantee. Section 4 contains extensive numerical studies and Section 5 demonstrates the performance of P-DCov via two real data sets. We conclude the paper with a short discussion in Section 6. Several technical lemmas and all proofs are relegated to the appendix.

2. Methods

First, we introduce some notations. For a random vector \mathbf{z} , $\|\mathbf{z}\|$ and $\|\mathbf{z}\|_1$ represent its Euclidean norm and ℓ_1 norm, respectively. A collection of n i.i.d. observations of \mathbf{z} is denoted as $\{\mathbf{z}_1, \dots, \mathbf{z}_n\}$, where $\mathbf{z}_k = (z_k^{(1)}, \dots, z_k^{(d)})^T$ represents the k -th observation. For any matrix \mathbf{M} , $\|\mathbf{M}\|_F$, $\|\mathbf{M}\|$ and $\|\mathbf{M}\|_{\max}$ denote its Frobenius norm, operator norm and max norm, respectively. $\|\mathbf{M}\|_{a,b}$ is the (a, b) norm defined as the ℓ_b norm of the vector consisting of column-wise ℓ_a norm of \mathbf{M} .

2.1. A brief review of distance covariance

As an important tool, distance covariance is briefly reviewed in this section with further details available in Székely *et al.* (2007). We introduce several definitions as follows.

DEFINITION 1. (*w-weighted L_2 norm*) Let $c_d = \frac{\pi^{(d+1)/2}}{\Gamma((d+1)/2)}$, for any positive integer d , where Γ is the Gamma function. Then for function γ defined on $\mathbb{R}^p \times \mathbb{R}^q$, the w -weighted L_2 norm of γ is defined by

$$\|\gamma(\boldsymbol{\tau}, \boldsymbol{\rho})\|_w^2 = \int_{\mathbb{R}^{p+q}} |\gamma(\boldsymbol{\tau}, \boldsymbol{\rho})|^2 w(\boldsymbol{\tau}, \boldsymbol{\rho}) d\boldsymbol{\tau} d\boldsymbol{\rho}, \quad \text{where } w(\boldsymbol{\tau}, \boldsymbol{\rho}) = (c_p c_q \|\boldsymbol{\tau}\|^{1+p} \|\boldsymbol{\rho}\|^{1+q})^{-1}.$$

DEFINITION 2. (*Distance covariance*) The distance covariance between random vectors $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^q$ with finite first moments is the nonnegative number $\mathcal{V}(\mathbf{x}, \mathbf{y})$ defined by

$$\mathcal{V}^2(\mathbf{x}, \mathbf{y}) = \|g_{x,y}(\boldsymbol{\tau}, \boldsymbol{\rho}) - g_x(\boldsymbol{\tau})g_y(\boldsymbol{\rho})\|_w^2,$$

where g_x , g_y and $g_{x,y}$ represent the characteristic functions of \mathbf{x} , \mathbf{y} and the joint characteristic function of \mathbf{x} and \mathbf{y} , respectively.

Suppose we observe random sample $\{(\mathbf{x}_k, \mathbf{y}_k) : k = 1, \dots, n\}$ from the joint distribution of (\mathbf{x}, \mathbf{y}) . We denote $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ and $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)$.

DEFINITION 3. (*Empirical distance covariance*) The empirical distance covariance between samples \mathbf{X} and \mathbf{Y} is the nonnegative random variable $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ defined by

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = S_1(\mathbf{X}, \mathbf{Y}) + S_2(\mathbf{X}, \mathbf{Y}) - 2S_3(\mathbf{X}, \mathbf{Y}),$$

where

$$\begin{aligned} S_1(\mathbf{X}, \mathbf{Y}) &= \frac{1}{n^2} \sum_{k,l=1}^n \|\mathbf{x}_k - \mathbf{x}_l\| \|\mathbf{y}_k - \mathbf{y}_l\|, \quad S_2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^n \|\mathbf{x}_k - \mathbf{x}_l\| \frac{1}{n^2} \sum_{k,l=1}^n \|\mathbf{y}_k - \mathbf{y}_l\|, \\ S_3(\mathbf{X}, \mathbf{Y}) &= \frac{1}{n^3} \sum_{k=1}^n \sum_{l,m=1}^n \|\mathbf{x}_k - \mathbf{x}_l\| \|\mathbf{y}_k - \mathbf{y}_m\|. \end{aligned}$$

With above definitions, Lemma 1 depicts the consistency of $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ as an estimator of $\mathcal{V}(\mathbf{x}, \mathbf{y})$. Lemma 2 shows the asymptotic distribution of $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ under the null hypothesis that \mathbf{x} and \mathbf{y} are independent. Corollary 1 reveals properties of the test statistic $n\mathcal{V}_n^2/S_2$ proposed in Székely *et al.* (2007).

LEMMA 1. (*Theorem 2 in Székely et al. (2007)*) Assume that $\mathbb{E}(\|\mathbf{x}\| + \|\mathbf{y}\|) < \infty$, then almost surely

$$\lim_{n \rightarrow \infty} \mathcal{V}_n(\mathbf{X}, \mathbf{Y}) = \mathcal{V}(\mathbf{x}, \mathbf{y}).$$

LEMMA 2. (*Theorem 5 in Székely et al. (2007)*) Assume that \mathbf{x} and \mathbf{y} are independent, and $\mathbb{E}(\|\mathbf{x}\| + \|\mathbf{y}\|) < \infty$, then as $n \rightarrow \infty$,

$$n\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) \xrightarrow{D} \|\zeta(\boldsymbol{\tau}, \boldsymbol{\rho})\|_w^2,$$

where \xrightarrow{D} represents convergence in distribution and $\zeta(\cdot, \cdot)$ denotes a complex-valued centered Gaussian random process with covariance function

$$R(\mathbf{u}, \mathbf{u}_0) = (g_x(\boldsymbol{\tau} - \boldsymbol{\tau}_0) - g_x(\boldsymbol{\tau})\overline{g_x(\boldsymbol{\tau}_0)})(g_y(\boldsymbol{\rho} - \boldsymbol{\rho}_0) - g_y(\boldsymbol{\rho})\overline{g_y(\boldsymbol{\rho}_0)}),$$

in which $\mathbf{u} = (\boldsymbol{\tau}, \boldsymbol{\rho})$, $\mathbf{u}_0 = (\boldsymbol{\tau}_0, \boldsymbol{\rho}_0)$.

COROLLARY 1. (Corollary 2 in Székely et al. (2007)) Assume that $\mathbb{E}(\|\mathbf{x}\| + \|\mathbf{y}\|) < \infty$.

- (i) If \mathbf{x} and \mathbf{y} are independent, then as $n \rightarrow \infty$, $n\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})/S_2 \xrightarrow{D} Q$ with $Q \stackrel{D}{=} \sum_{j=1}^{\infty} \lambda_j Z_j^2$, where $Z_j \stackrel{i.i.d}{\sim} \mathcal{N}(0, 1)$ and $\{\lambda_j\}$ are non-negative constants depending on the distribution of (\mathbf{x}, \mathbf{y}) ; $\mathbb{E}(Q) = 1$.
- (ii) If \mathbf{x} and \mathbf{y} are dependent, then as $n \rightarrow \infty$, $n\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})/S_2 \xrightarrow{P} \infty$.

2.2. Conditional independence test via projected distance covariance (P-DCov)

Here, we consider the case where \mathbf{G}_x and \mathbf{G}_y are linear in (1), which leads to the following factor model setup:

$$\mathbf{x} = \mathbf{B}_x \mathbf{f} + \boldsymbol{\epsilon}_x, \quad \mathbf{y} = \mathbf{B}_y \mathbf{f} + \boldsymbol{\epsilon}_y, \quad (4)$$

where \mathbf{B}_x and \mathbf{B}_y are factor loading matrices of dimension $p \times K$ and $q \times K$ respectively, and \mathbf{f} is the K -dimensional vector of common factors. Here, the number of common factors K could grow to infinity and the matrices \mathbf{B}_x and \mathbf{B}_y are assumed to be sparse to reflect that \mathbf{x} and \mathbf{y} only depend on several important factors. As a result, we will impose regularization on the estimation of \mathbf{B}_x and \mathbf{B}_y . Now, we are in the position to propose a test for problem (2). We first provide an estimate for the idiosyncratic components $\boldsymbol{\epsilon}_x$ and $\boldsymbol{\epsilon}_y$, and then calculate distance covariance between the estimates. More generally, we project \mathbf{x} and \mathbf{y} onto the space orthogonal to the linear space spanned by \mathbf{f} and evaluate the dependency between the projected vectors. The conditional independence test is summarized in the following steps.

Step 1: Estimate factor loading matrices \mathbf{B}_x and \mathbf{B}_y by the penalized least square (PLS) estimators $\hat{\mathbf{B}}_x$ and $\hat{\mathbf{B}}_y$ defined as follows.

$$\hat{\mathbf{B}}_x = \arg \min_{\mathbf{B}} \frac{1}{2} \|\mathbf{X} - \mathbf{B}\mathbf{F}\|_F^2 + \sum_{j,k} p_{\lambda_1}(|B_{jk}|), \quad (5)$$

$$\hat{\mathbf{B}}_y = \arg \min_{\mathbf{B}} \frac{1}{2} \|\mathbf{Y} - \mathbf{B}\mathbf{F}\|_F^2 + \sum_{j,k} p_{\lambda_2}(|B_{jk}|), \quad (6)$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)$, $\mathbf{F} = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n)$, $p_{\lambda}(\cdot)$ is the penalty function with penalty level λ .

Step 2: Estimate the error vectors $\boldsymbol{\epsilon}_{i,x}$ and $\boldsymbol{\epsilon}_{i,y}$ by

$$\begin{aligned} \hat{\boldsymbol{\epsilon}}_{i,x} &= \mathbf{x}_i - \hat{\mathbf{B}}_x \mathbf{f}_i = (\mathbf{B}_x - \hat{\mathbf{B}}_x) \mathbf{f}_i + \boldsymbol{\epsilon}_{i,x}, \\ \hat{\boldsymbol{\epsilon}}_{i,y} &= \mathbf{y}_i - \hat{\mathbf{B}}_y \mathbf{f}_i = (\mathbf{B}_y - \hat{\mathbf{B}}_y) \mathbf{f}_i + \boldsymbol{\epsilon}_{i,y}, \quad i = 1, \dots, n. \end{aligned}$$

Step 3: Define the estimated error matrices $\hat{\mathbf{E}}_x = (\hat{\boldsymbol{\epsilon}}_{1,x}, \dots, \hat{\boldsymbol{\epsilon}}_{n,x})$ and $\hat{\mathbf{E}}_y = (\hat{\boldsymbol{\epsilon}}_{1,y}, \dots, \hat{\boldsymbol{\epsilon}}_{n,y})$. Calculate the empirical distance covariance between $\hat{\mathbf{E}}_x$ and $\hat{\mathbf{E}}_y$ as

$$\mathcal{V}_n^2(\hat{\mathbf{E}}_x, \hat{\mathbf{E}}_y) = S_1(\hat{\mathbf{E}}_x, \hat{\mathbf{E}}_y) + S_2(\hat{\mathbf{E}}_x, \hat{\mathbf{E}}_y) - 2S_3(\hat{\mathbf{E}}_x, \hat{\mathbf{E}}_y).$$

Step 4: Define the P-DCov test statistic as $T(\mathbf{x}, \mathbf{y}, \mathbf{f}) = n\mathcal{V}_n^2(\hat{\mathbf{E}}_x, \hat{\mathbf{E}}_y)/S_2(\hat{\mathbf{E}}_x, \hat{\mathbf{E}}_y)$.

Step 5: With predetermined significance level α , we reject the null hypothesis when $T(\mathbf{x}, \mathbf{y}, \mathbf{f}) > (\Phi^{-1}(1 - \alpha/2))^2$.

Theoretical properties of the proposed conditional independence test will be studied in Section 3. In the above method, we implicitly assume that the number of variables K is large so that the penalized least-squares methods are used. When the number of variables K is small, we can take $\lambda_1 = \lambda_2 = 0$ so that no penalization is imposed.

2.3. Building graphs via conditional independence test

Now we explore a specific application of our conditional independence test to graphical models. To identify the conditional independence relationship in a graphical model, i.e., $z^{(i)} \perp\!\!\!\perp z^{(j)} | \mathbf{z} \setminus \{z^{(i)}, z^{(j)}\}$, we assume

$$z_k^{(i)} = \beta_{1,ij}^\top \mathbf{f}_k + \epsilon_k^{(i)}, \quad z_k^{(j)} = \beta_{2,ij}^\top \mathbf{f}_k + \epsilon_k^{(j)}, \quad k = 1, \dots, n, \quad (7)$$

where $\mathbf{f}_k = (\mathbf{z}_k^{(-i,-j)})^\top$ represents all coordinates of \mathbf{z}_k other than $\mathbf{z}_k^{(i)}$ and $\mathbf{z}_k^{(j)}$, and $\beta_{1,ij}$ and $\beta_{2,ij}$ are $d - 2$ dimensional regression coefficients. Under model (7), we decide whether edge e_{ij} will be drawn through directly testing $z^{(i)} \perp\!\!\!\perp z^{(j)} | \mathcal{L}(\mathbf{z}^{(-i,-j)})$, where $\mathcal{L}(\mathbf{f})$ is the linear space spanned by \mathbf{f} .

More specifically, for each node pair $\{(i, j) : 1 \leq i < j \leq d\}$, we define $T^{(i,j)} = T(z^{(i)}, z^{(j)}, \mathbf{z}^{(-i,-j)})$ using the same steps as in Section 2.2 as the test for the current null hypothesis:

$$H_{0,ij} : \epsilon^{(i)} \perp\!\!\!\perp \epsilon^{(j)}. \quad (8)$$

We now summarize the testing results by a graph in which nodes represent variables in \mathbf{z} and the edge e_{ij} between node i and node j is drawn only when $H_{0,ij}$ is rejected at level α .

In (7), the factors are created internally via the observations on remaining nodes $\mathbf{z} \setminus \{z^{(i)}, z^{(j)}\}$. In financial applications, it is often desirable to build graphs when conditioning on external factors. In such cases, it is straightforward to change the factors in (7) to external factors.

We will demonstrate the two different types of conditional dependency graphs via examples in Sections 4 and 5.

2.4. Graph estimation with FDR control

Through the graph building process described in Section 2.3, we can carry out $\bar{d} = d(d-1)/2$ P-DCov tests simultaneously and we wish to control the *false discovery rate* (FDR) at a pre-specified level $0 < \alpha < 1$. Let R_F and R be the number of falsely rejected hypotheses and the number of total rejections, respectively. The *false discovery proportion* (FDP) is defined as $R_F / \max\{1, R\}$ and the FDR is the expectation of FDP.

In the literature, various procedures have been proposed for conducting large-scale multiple hypothesis testing via FDR control. In this work, we will follow the most commonly used Benjamini and Hochberg (BH) procedure developed in the seminal work of Benjamini and Hochberg (1995), where P-values of all marginal tests are compared. More specifically, let $P_{(1)} \leq P_{(2)} \leq \dots \leq P_{(\bar{d})}$ be the ordered P-values of the \bar{d} hypotheses given in (8). Let $s = \max\{0 \leq i \leq \bar{d} : P_{(i)} \leq \alpha i / \bar{d}\}$, and we reject the s hypotheses $H_{0,ij}$ with the smallest P-values. We will demonstrate the performance of this strategy via real data examples.

2.5. Extension to functional projection

In the P-DCov described in Section 2.2, we assume the conditional dependency of \mathbf{x} and \mathbf{y} given factor \mathbf{f} is expressed via a linear form of \mathbf{f} . In other words, we are projecting \mathbf{x} and \mathbf{y} onto the space orthogonal to $\mathcal{L}(\mathbf{f})$ and evaluate the dependence between the projected vectors. Although this linear projection assumption makes the theoretical development easier and delivers the main idea of this work, a natural extension is to consider a nonlinear projection. In particular, we consider the following additive generalization (Stone, 1985) of the factor model setup:

$$\mathbf{x} = \sum_{j=1}^K \mathbf{g}_j^x(f_j) + \epsilon_x, \quad \mathbf{y} = \sum_{j=1}^K \mathbf{g}_j^y(f_j) + \epsilon_y, \quad (9)$$

where $\{\mathbf{g}_j^x(\cdot), \mathbf{g}_j^y(\cdot), j = 1, \dots, K\}$ are unknown vector-valued functions we would like to estimate. In (9), we consider the additive space spanned by factor \mathbf{f} . By this extension, we could identify more general conditional dependency structures between \mathbf{x} and \mathbf{y} given \mathbf{f} . This is a special case of (1), but avoids the issue of curse of dimensionality.

In the high-dimensional setup where K is large, we can use a penalized additive model (Ravikumar *et al.*, 2009; Fan *et al.*, 2011) to estimate the unknown functions. The conditional independence test described in Section 2.2 could be modified by replacing the linear regression with the (penalized) additive model regression. We will investigate the P-DCov method coupled with the sparse additive model (Ravikumar *et al.*, 2009) in numerical studies.

3. Theoretical Results

In this section, we derive the asymptotic properties of our conditional independence test. First, we introduce several assumptions on ϵ_x , ϵ_y and \mathbf{f} .

CONDITION 1. $\mathbb{E}\epsilon_x = \mathbb{E}\epsilon_y = \mathbf{0}$, $\mathbb{E}\|\epsilon_x\| < \infty$, $\mathbb{E}\|\epsilon_y\| < \infty$, $\mathbb{E}\|\epsilon_x\|^2 < \infty$, $\mathbb{E}\|\epsilon_y\|^2 < \infty$.

CONDITION 2. We denote h_x as the density function of random variable x . Let us assume that the densities of $\|\epsilon_{1,x} - \epsilon_{2,x}\|$ and $\|\epsilon_{1,y} - \epsilon_{2,y}\|$ are bounded on $[0, C_0]$, for some positive constant C_0 . In other words, there exists a positive constant M ,

$$\max_{t \in [0, C_0]} h_{\|\epsilon_{i,x} - \epsilon_{j,x}\|}(t) \leq M, \quad \max_{t \in [0, C_0]} h_{\|\epsilon_{i,y} - \epsilon_{j,y}\|}(t) \leq M.$$

Conditions 1 and 2 impose mild moment and distributional assumptions on random errors ϵ_x and ϵ_y . To better understand when the proposed method works, we give the following high-level assumption, whose justifications are noted below.

CONDITION 3. There exist constants $C_1 > 1$ and $\gamma > 0$, such that for any $C_2 > 1$, with probability greater than $1 - C_1^{-C_2}$, we have for any n ,

$$\|(\mathbf{B}_x - \hat{\mathbf{B}}_x)\mathbf{F}\|_{2,\infty} \leq C_2 a_n, \quad \|(\mathbf{B}_y - \hat{\mathbf{B}}_y)\mathbf{F}\|_{2,\infty} \leq C_2 a_n,$$

where the sequence $a_n = o\{(n^{(1+\gamma)} \log n)^{-1/3}\}$.

CONDITION 4. Let $\mathbf{B}_{x,l}$ denote the l -th row of \mathbf{B}_x , and similarly we define $\hat{\mathbf{B}}_{x,l}$, $\mathbf{B}_{y,l}$ and $\hat{\mathbf{B}}_{y,l}$. We assume for any fixed l ,

$$\|\mathbf{B}_{x,l} - \hat{\mathbf{B}}_{x,l}\|_1 = O_p(e_n), \quad \|\mathbf{B}_{y,l} - \hat{\mathbf{B}}_{y,l}\|_1 = O_p(e_n),$$

where sequences e_n and a_n in Condition 3 satisfy $a_n e_n = o(\frac{1}{\sqrt{n \log K}})$.

REMARK 1. Conditions 3 and 4 are mild conditions that are imposed to ensure the quality of the projection and guarantee the theoretical properties regarding our conditional independence test. For example, one could directly call the results from penalized least squares for high-dimensional regression (Bühlmann and Van De Geer, 2011; Hastie *et al.*, 2015) and robust estimation (Belloni *et al.*, 2011; Wang, 2013; Fan *et al.*, 2016b). We now discuss two special examples as follows.

- (i) (K is fixed) In this fixed dimensional case, it is straightforward to verify that the projection based on ordinary least squares satisfies the two conditions.

(ii) (*Sparse Linear Projection*) Let $\mathbf{B}_x = (\mathbf{b}_1^T, \mathbf{b}_2^T, \dots, \mathbf{b}_p^T)^T$ and $\widehat{\mathbf{B}}_x = (\widehat{\mathbf{b}}_1^T, \widehat{\mathbf{b}}_2^T, \dots, \widehat{\mathbf{b}}_p^T)^T$. Note that the graphical model case corresponds to $p = 1$. We apply the popular L_1 -regularized least squares for each dimension of \mathbf{x} regressing on the factor \mathbf{F} . Here, we further assume the true regression coefficient \mathbf{b}_j is sparse for each j with $S_j = \{k : (\mathbf{b}_j)_k \neq 0\}$, $\hat{S}_j = \{k : (\widehat{\mathbf{b}}_j)_k \neq 0\}$ and $|S_j| = s_j$. From Theorem 11.1, Example 11.1 and Theorem 11.3 in Hastie et al. (2015), and since $\{\mathbf{f}_i\}_{i=1}^n$ are i.i.d., we have with high probability, $\|\widehat{\mathbf{b}}_j - \mathbf{b}_j\| \leq C\sqrt{\frac{s_j \log K}{n}}$, $\hat{S}_j = S_j$ and $\max_i \|(\mathbf{f}_i)_{S_j}\| \leq s_j \log n$. Then, we have with high probability, for each $i = 1, \dots, n$ and $j = 1, \dots, p$,

$$\|(\widehat{\mathbf{b}}_j - \mathbf{b}_j)^T \mathbf{f}_i\| = \|(\widehat{\mathbf{b}}_j - \mathbf{b}_j)_{S_j}^T (\mathbf{f}_i)_{S_j}\| \leq \|(\widehat{\mathbf{b}}_j - \mathbf{b}_j)_{S_j}\| \|(\mathbf{f}_i)_{S_j}\| \leq C s_{\max} \log n \sqrt{\frac{s_{\max} \log K}{n}}, \quad (10)$$

where $s_{\max} = \max_j s_j$. It is now easy to verify that Condition 3 and 4 are satisfied even under the ultra-high-dimensional case where $\log K = o(n^a)$, $0 < a < 1/3$. We would like to omit the details here for brevity about the specification of various constants.

THEOREM 1. Under Conditions 1 and 3,

$$\mathcal{V}_n^2(\hat{\epsilon}_x, \hat{\epsilon}_y) \xrightarrow{P} \mathcal{V}^2(\epsilon_x, \epsilon_y).$$

In particular, when ϵ_x and ϵ_y are independent, $\mathcal{V}_n^2(\hat{\epsilon}_x, \hat{\epsilon}_y) \xrightarrow{P} 0$.

Theorem 1 shows that the sample distance covariance between the estimated residual vectors converges to the distance covariance between the population error vectors. It enables us to use the distance covariance of the estimated residual vectors to construct the conditional independence test as described in Section 2.2.

THEOREM 2. Under Conditions 1-4, and the null hypothesis that $\epsilon_x \perp \epsilon_y$ (or equivalently $\mathbf{x} \perp \mathbf{y} | \mathbf{f}$),

$$n\mathcal{V}_n^2(\hat{\epsilon}_x, \hat{\epsilon}_y) \xrightarrow{\mathcal{D}} \|\zeta\|^2,$$

where ζ is a zero-mean Gaussian process defined analogously as in Lemma 2.

Theorem 2 provides the asymptotic distribution of the test statistic $T(\mathbf{x}, \mathbf{y}, \mathbf{f})$ under the null hypothesis, which is the basis of Theorem 3.

COROLLARY 2. Under the same conditions of Theorem 2,

$$n\mathcal{V}_n^2(\hat{\epsilon}_x, \hat{\epsilon}_y)/S_2(\hat{\epsilon}_x, \hat{\epsilon}_y) \xrightarrow{\mathcal{D}} Q, \quad \text{where } Q \stackrel{\mathcal{D}}{=} \sum_{j=1}^{\infty} \lambda_j Z_j^2,$$

where $Z_j \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $\{\lambda_j\}$ are non-negative constants depending on the distribution of (\mathbf{x}, \mathbf{y}) ; $\mathbb{E}(Q) = 1$.

THEOREM 3. Consider the test that rejects conditional independence when

$$\frac{n\mathcal{V}_n^2(\hat{\epsilon}_x, \hat{\epsilon}_y)}{S_2} > (\Phi^{-1}(1 - \alpha/2))^2, \quad (11)$$

where $\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0, 1)$. Let $\alpha_n(\mathbf{x}, \mathbf{y}, \mathbf{f})$ denote its associated type I error. Then under Conditions 1-4, for all $0 < \alpha \leq 0.215$,

$$(i) \lim_{n \rightarrow \infty} \alpha_n(\mathbf{x}, \mathbf{y}, \mathbf{f}) \leq \alpha,$$

$$(ii) \sup_{\epsilon_x \perp \epsilon_y} \lim_{n \rightarrow \infty} \alpha_n(\mathbf{x}, \mathbf{y}, \mathbf{f}) = \alpha.$$

Part (i) of Theorem 3 indicates the proposed test with critical region (11) has an asymptotic significance error at most α . Part (ii) of Theorem 3 implies that there exists a pair (ϵ_x, ϵ_y) such that the pre-specified significant level α is achieved asymptotically. In other words, the size of testing $H_0 : \epsilon_x \perp \epsilon_y$ is α .

REMARK 2. When the sample size n is small, the theoretical critical value in (11) could sometimes be too conservative in practice (Székely et al., 2007). Therefore, we recommend using random permutation to get a reference distribution for the test statistic $T(\mathbf{x}, \mathbf{y}, \mathbf{f})$ under H_0 . Random permutation is used to decouple $\epsilon_{i,x}$ and $\epsilon_{i,y}$ so that the resulting pair $(\epsilon_{\pi(i),x}, \epsilon_{i,y})$ follows the null model, where $\{\pi(1), \dots, \pi(n)\}$ are a random permutation of indices $\{1, \dots, n\}$. Here, we set the number of permutations $R(n) = \lfloor 200 + 5000/n \rfloor$ as in Székely et al. (2007). Consequently, we can also estimate the P-value associated with the conditional independence test based on the quantiles of the test statistics over $R(n)$ random permutations.

4. Monte Carlo Experiments

In this section, we investigate the performance of P-DCov with five simulation examples. In Example 4.1, we consider a factor model and test the conditional independence between two vectors \mathbf{x} and \mathbf{y} given their common factor \mathbf{f} , via P-DCov. In Examples 4.2, we investigate the classical Gaussian graphical model. In Example 4.3, we consider the case of general graphical model without the Gaussian assumption. In Examples 4.4 and 4.5, we consider the case of factor based dependency graph and a general graphical model with external factors, respectively.

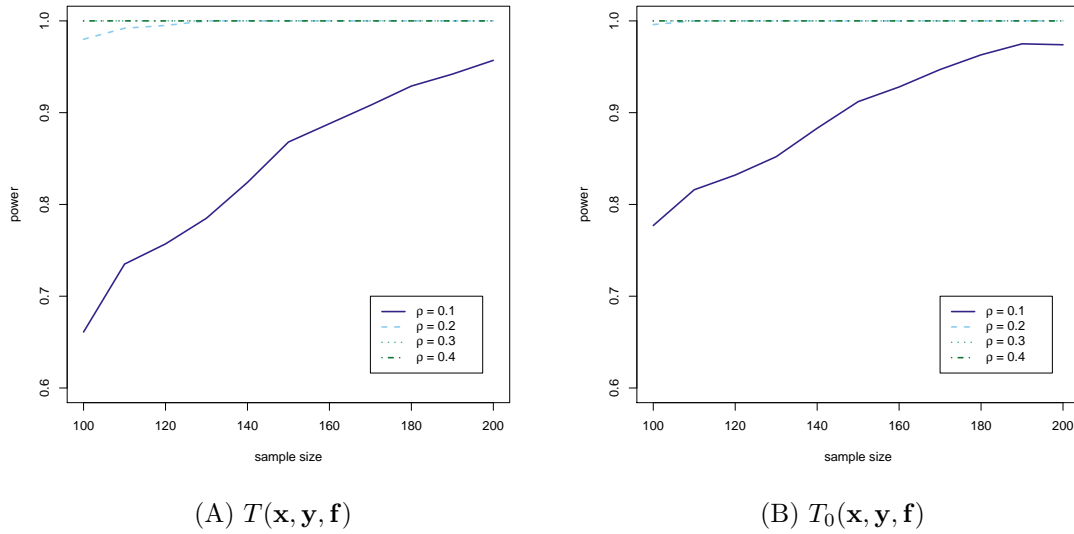
EXAMPLE 4.1. [High-dimensional factor model] Let $p = 5$, $q = 10$ and $K = 1000$. Assume the rows of \mathbf{B}_x and rows of \mathbf{B}_y are i.i.d. distributed as $\mathbf{z}_K = (\mathbf{z}_1^T, \mathbf{z}_2^T)^T$, where \mathbf{z}_1 is a 3-dimensional vector with elements i.i.d. from $Unif[2, 3]$ and $\mathbf{z}_2 = \mathbf{0}_{K-3}$. $\{\mathbf{f}_i\}_{i=1}^n$ are i.i.d. from $\mathcal{N}(0, I_K)$. We generate n i.i.d. copies $\{\mathbf{r}_i\}_{i=1}^n$ from log-normal distribution $\ln \mathcal{N}(0, \Sigma)$ where Σ is an equal correlation matrix of size $(p+q) \times (p+q)$ with $\Sigma_{jk} = \rho$ when $j \neq k$ and $\Sigma_{jj} = 1$. $\epsilon_{i,x}$ and $\epsilon_{i,y}$ are the centered version of the first p coordinates and the last q coordinates of \mathbf{r}_i . Then, $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$ are generated according to $\mathbf{x}_i = \mathbf{B}_x \mathbf{f}_i + \epsilon_{i,x}$ and $\mathbf{y}_i = \mathbf{B}_y \mathbf{f}_i + \epsilon_{i,y}$ correspondingly.

In Example 4.1, we consider a high-dimensional factor model with sparsity structure. Note that the errors are generated from a heavy tail distribution to demonstrate the proposed test works beyond Gaussian errors. We assume each coordinate of \mathbf{x} and \mathbf{y} only depends on the first three factors. We calculate $T(\mathbf{x}, \mathbf{y}, \mathbf{f})$ in the P-DCov test, and $T_0(\mathbf{x}, \mathbf{y}, \mathbf{f})$ in which we replace $\hat{\epsilon}_{i,x}$ and $\hat{\epsilon}_{i,y}$ by the true $\epsilon_{i,x}$ and $\epsilon_{i,y}$ as an oracle test to compare with. To get reference distributions of $T(\mathbf{x}, \mathbf{y}, \mathbf{f})$ and $T_0(\mathbf{x}, \mathbf{y}, \mathbf{f})$, we follow the permutation procedure as described in Section 3. In this example, we set the significance level $\alpha = 0.1$. We vary the sample size from 100 to 200 with increment of 10 and show the empirical power based on 1000 repetitions for both $T(\mathbf{x}, \mathbf{y}, \mathbf{f})$ and $T_0(\mathbf{x}, \mathbf{y}, \mathbf{f})$ in Figure 1 for $\rho \in \{0.1, 0.2, 0.3, 0.4\}$. In the implementation of penalized least squares in Step 1, we use R package `glmnet` with the default tuning parameter selection method (10-fold cross-validation) and perform least square on the selected variables to reduce estimation bias of these estimated parameters.

From Figure 1, it is clear that as the sample size or ρ increases, the empirical power also increases in general. Also, comparing the panels (A) and (B) in Figure 1, we see that when the sample size is small, the P-DCov test has smaller power than the oracle test, however, the difference between them becomes

Table 1. Type I error of Example 1

Test based on $\hat{\epsilon}_x$ and $\hat{\epsilon}_y$											
n	100	110	120	130	140	150	160	170	180	190	200
	0.129	0.112	0.117	0.095	0.104	0.108	0.098	0.113	0.102	0.117	0.111
Test based on ϵ_x and ϵ_y											
n	100	110	120	130	140	150	160	170	180	190	200
	0.089	0.092	0.078	0.112	0.104	0.090	0.099	0.113	0.104	0.106	0.096

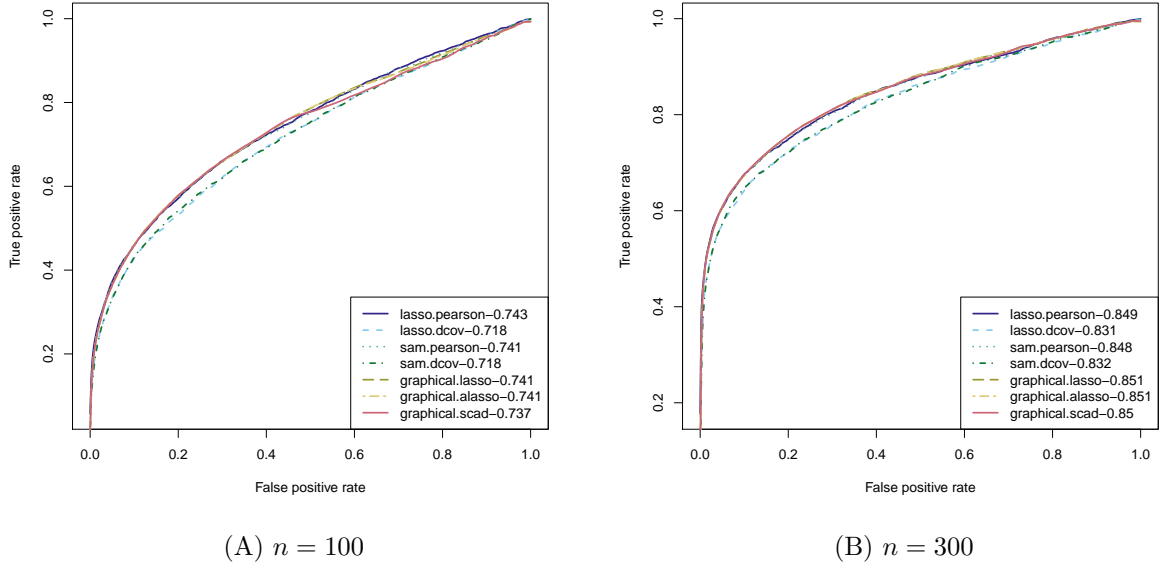
Fig. 1. Power-sample size graph of Example 1

negligible as the sample size increases. This is consistent with our theory regarding the asymptotic distribution of the test statistics. When $\rho = 0$, Table 1 reports the empirical type I error for both P-DCov as well as the oracle version. It is clear that the type I error of P-DCov is under good control as the sample size increases.

EXAMPLE 4.2. *[Gaussian graphical model]* We consider a standard Gaussian graphical model with precision matrix $\Omega = \Sigma^{-1}$, where Ω is a tridiagonal matrix of size $d \times d$, and is associated with the autoregressive process of order one. We set $d = 30$ and the (i, j) -element in Σ to be $\sigma_{i,j} = \exp(-|s_i - s_j|)$, where $s_1 < s_2 < \dots < s_d$. In addition,

$$s_i - s_{i-1} \stackrel{i.i.d}{\sim} \text{Uniform}(1, 3), \quad i = 2, \dots, d.$$

In this example, we would like to compare the proposed P-DCov with the state-of-the-art approaches for recovering Gaussian graphical models. In terms of recovering structure Ω , we compare lasso.dcov (projection by LASSO followed by distance covariance), sam.dcov (projection by sparse additive model followed by distance covariance), lasso.pearson (projection by LASSO followed by Pearson correlation), sam.pearson (projection by sparse additive model followed by Pearson correlation) with three popular estimators corresponding to the LASSO, adaptive LASSO and SCAD penalized likelihoods (called graphical.lasso, graphical.alasso and graphical.scad on the graph) for the precision matrix (Friedman *et al.*, 2008; Fan *et al.*, 2009). Here, lasso.dcov and sam.dcov are two examples of our P-DCov methods. We use R package SAM to fit the sparse additive model. To evaluate the performances, we construct receiver

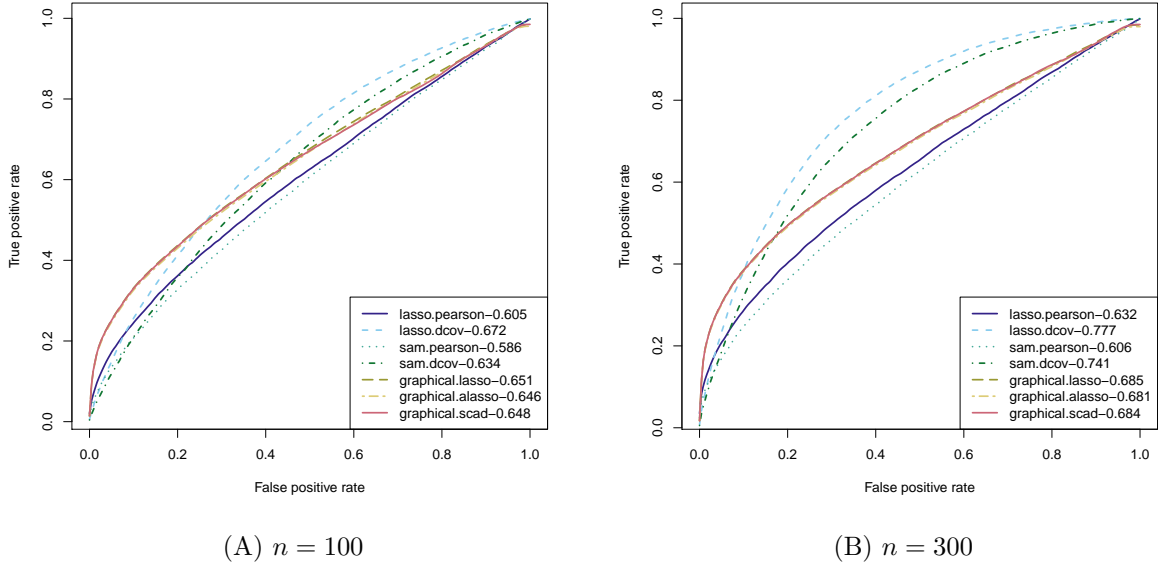
Fig. 2. ROC curves for Gaussian graphical models

operating characteristic (ROC) curves for each method with sample sizes $n = 100$ and $n = 300$. The process of constructing the ROC curves involves conducting the P-DCov test for each pair of nodes and record the corresponding P-values. In each of the ROC curve, true positive rates (TPR) are plotted against false positive rates (FPR) at various thresholds of those P-values (“TP” means the true entry of the precision matrix is nonzero and estimated as nonzero; “FP” means the true entry of the precision matrix is zero but estimated as nonzero.) We follow the implementation in Fan *et al.* (2009) for the three penalized likelihood estimators. The average results over 100 replications of different methods are reported in Figure 2. The associated AUC (Area Under the Curve) for each method is also displayed in the legend of the figure.

We observe that lasso.pearson and sam.pearson perform similarly with the penalized likelihood methods when $n = 100$. On the other hand, lasso.dcov and sam.dcov lead to slightly smaller AUC value due to the use of the distance covariance, which is expected for the Gaussian model. This shows that we do not pay a big price for using the more complicated distance covariance and sparse additive model.

EXAMPLE 4.3. *[A general graphical model] We consider a general graphical model with a combination of multivariate t distribution and multivariate Gaussian distribution. The dimension of \mathbf{x} is $d = 30$. In detail, $\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T)^T$ where \mathbf{x}_1 follows a 20 dimensional multivariate t distribution with degrees of freedom 5, location parameter 0 and identity covariance matrix and \mathbf{x}_2 follows the same Gaussian graphical model as in Example 4.2 except the dimension is now 10. In addition, \mathbf{x}_1 and \mathbf{x}_2 are independent.*

To generate a multivariate t -distribution, we first generate a random vector \mathbf{w}_{20} from the standard multivariate Gaussian distribution and an independent random variable $\tau \sim \chi^2(5)$ and then set $\mathbf{x}_1 = \mathbf{w}/\sqrt{\tau}$. One important fact about the multivariate t distribution is that the zero element in the precision matrix does not imply conditional independence like the case of Gaussian graphical models (Finagold and Drton, 2009). Indeed, for \mathbf{x}_1 , we actually have the fact that $\mathbf{x}_1^{(i)}$ and $\mathbf{x}_1^{(j)}$ are dependent given $\mathbf{x}_1^{(-i,-j)}$ for any pair $1 \leq i \neq j \leq 20$. Consequently, the Gaussian likelihood based methods will falsely claim that all the components of \mathbf{x}_1 are independent.

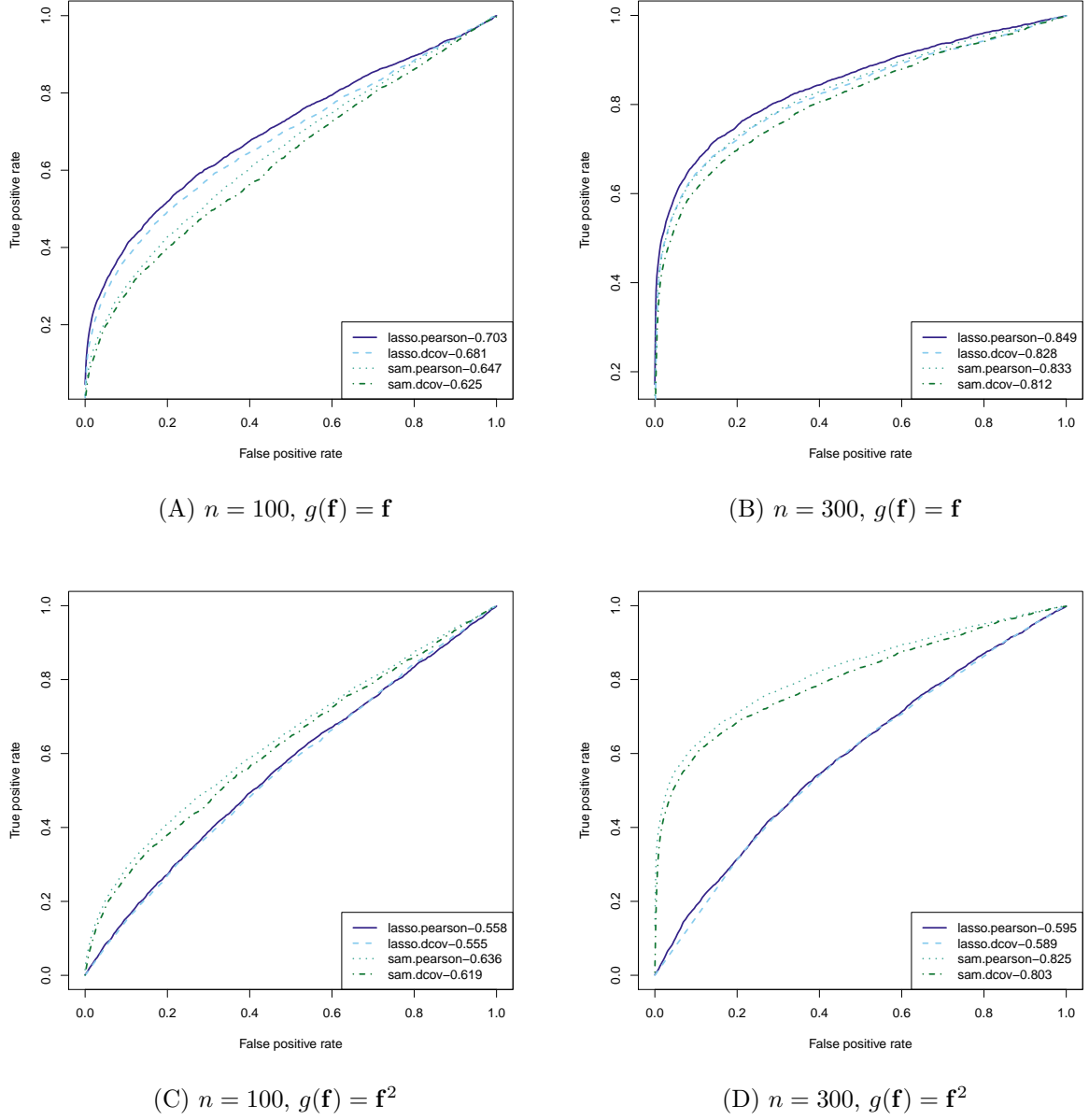
Fig. 3. ROC curves for a general graphical model

The average ROC curve results are rendered in Figure 3. As expected, by using the new projection based distance covariance method for testing conditional independence, lasso.dcov outperforms all the other methods in terms of AUC, with a more evident advantage when $n = 300$. One interesting observation is that: in the region where FPR is very low, the likelihood based methods actually outperform P-DCov methods. One possible reason is that the likelihood based methods are more capable of capturing the conditional dependency structure within \mathbf{x}_2 as it follows a Gaussian graphical model.

EXAMPLE 4.4. *[Factor based dependency graph]* We consider a dependency graph with the contribution of external factors. In particular, we generate $\mathbf{u} \sim N(0, \mathbf{\Omega})$, where $\mathbf{\Omega}$ is the same tridiagonal matrix used in Example 4.2 and $\mathbf{f} \sim N(0, \mathbf{I})$, then the observation $\mathbf{x} = \mathbf{u} + \mathbf{Q}g(\mathbf{f})$ where \mathbf{Q} is a sparse coefficient matrix that dictates how each dimension of \mathbf{x} depends on the factor $g(\mathbf{f})$. In particular, the generation of \mathbf{Q} follows the setting in Cai et al. (2013). For each element Q_{ij} , we first generate a Bernoulli distribution with success probability 0.2 to determine whether Q_{ij} is 0 or not. If Q_{ij} is not 0, we then generate $Q_{ij} \sim \text{Uniform}(0.5, 1)$. Here we consider two forms of $g(\cdot)$, namely $g(\mathbf{f}) = \mathbf{f}$ and $g(\mathbf{f}) = \mathbf{f}^2$.

We report results regarding the average ROC curves for lasso.pearson, lasso.dcov, sam.pearson and sam.dcov. The results for both $g(\mathbf{f}) = \mathbf{f}$ and $g(\mathbf{f}) = \mathbf{f}^2$ are depicted in Figure 4. Note that we are not building a conditional dependency graph among \mathbf{x} , but a dependency graph of \mathbf{x} conditioning on the external factor \mathbf{f} . There are some insightful observations from the figure. First of all, by looking at the first case when $g(\mathbf{f}) = \mathbf{f}$, it is clear that lasso.pearson is the best as it takes advantage of the sparse linear structure paired with the Gaussian distribution of the residual. By using the distance covariance as a dependency measure, or by using the sparse additive model as a projection method, it is reassuring that we do not lose much efficiency. Second, for the case when $g(\mathbf{f}) = \mathbf{f}^2$, we can see a substantial advantage of the sparse additive model based methods as it can capture this nonlinear contribution of the factors to the dependency structure of \mathbf{x} .

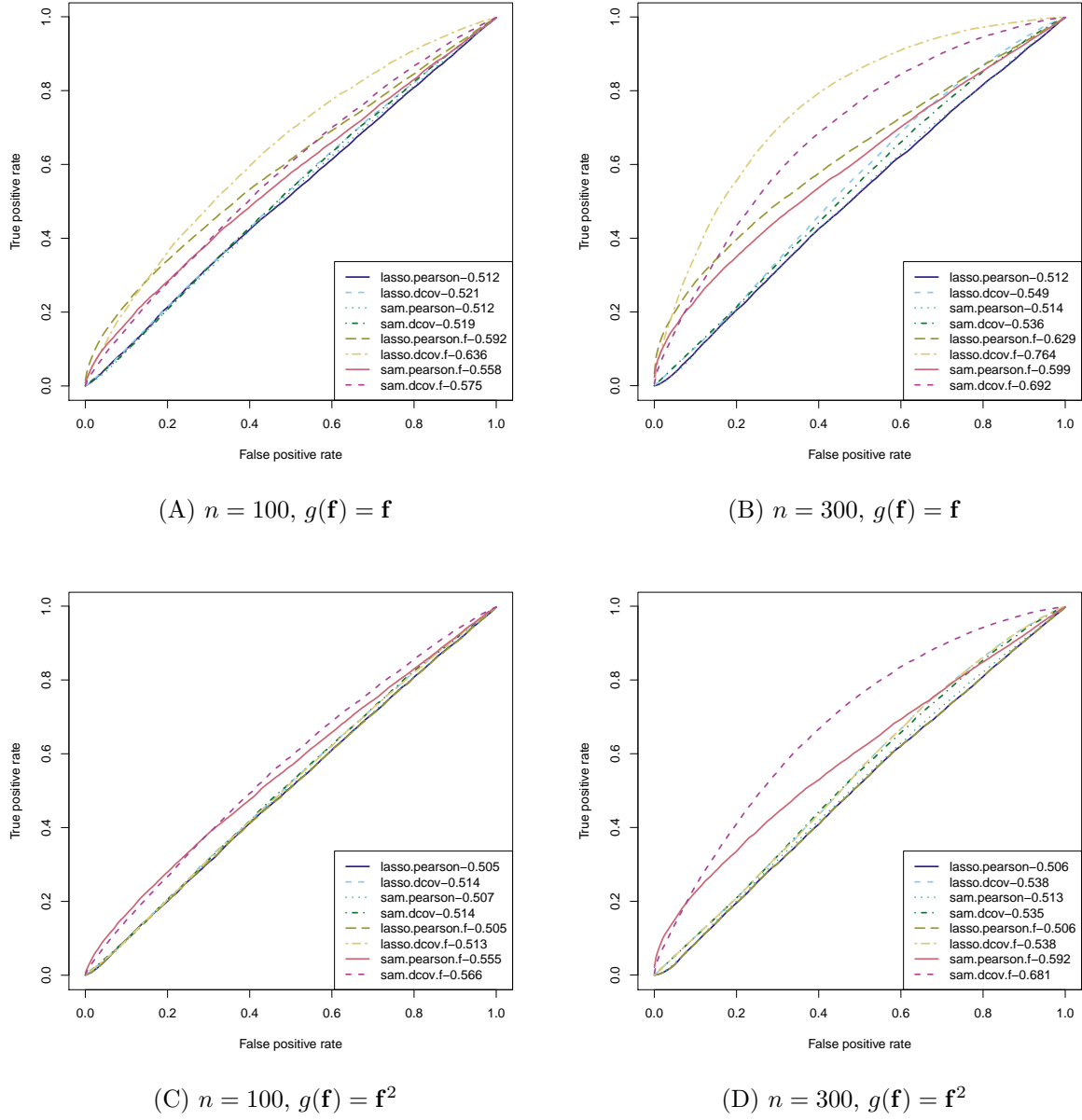
EXAMPLE 4.5. *[A general graphical model with external factors]* We consider a general conditional dependency graph with the contribution of external factors by combining the ingredients of Examples 4.3

Fig. 4. ROC curves for factor based dependency graph

and 4.4. In particular, we generate \mathbf{u} from Example 4.3 and $\mathbf{f} \sim N(0, I_{30})$, then set $\mathbf{x} = \mathbf{u} + \mathbf{Q}g(\mathbf{f})$ where \mathbf{Q} is the same as Example 4.4. We also consider $g(\mathbf{f}) = \mathbf{f}$ and $g(\mathbf{f}) = \mathbf{f}^2$.

In this example, we would like to investigate the performance of a two-step projection method. In particular, we first project \mathbf{x} onto the space spanned by \mathbf{f} and denote the residual by $\hat{\mathbf{u}}$. Then we explore the conditional dependency structure of $\hat{\mathbf{u}}^{(i)}$ and $\hat{\mathbf{u}}^{(j)}$ given $\hat{\mathbf{u}}^{(-i,-j)}$ by projecting them onto the space orthogonal to the space (linearly or additively) spanned by $\hat{\mathbf{u}}^{(-i,-j)}$. Here, we compare the performances of methods using the external factor and those that ignore them. The average ROC curves are rendered in Figure 5.

From the figure, we see that first of all, when $g(\mathbf{f}) = \mathbf{f}$, the methods using external factors outperform their counterparts without using the information with the best method being lasso.dcov (lasso regression

Fig. 5. ROC curves for a general graphical model with external factors

based projection coupled with distance covariance). Second, when we have nonlinear factors, using the factors do not necessarily help when we only consider linear projection. For example, the performances of `lasso.pearson` and `lasso.pearson.f` in panel (c) illustrates this point. On the other hand, by using sparse additive model based projection, we have a substantial gain over all the remaining methods especially for $n = 300$.

5. Real Data Analysis

5.1. Financial Data

In the first empirical example, we consider the Fama-French three-factor model (Fama and French, 1993). We collect daily excess returns of 90 stocks among the S&P 100 index, which are available between August

19, 2004 and August 19, 2005. We chose the starting date as Google’s Initial Public Offering date, and consider one year of daily excess returns since then. In particular, we consider the following three-factor model

$$r_{it} - r_{ft} = \beta_{i,\text{MKT}}(\text{MKT}_t - r_{ft}) + \beta_{i,\text{SMB}}\text{SMB}_t + \beta_{i,\text{HML}}\text{HML}_t + u_{it},$$

for $i = 1, \dots, 90$ and $t = 1, \dots, 252$. At time t , r_{it} represents the return for stock i , r_{ft} is the risk-free return rate, and MKT, SMB and HML constitute market, size and value factors.

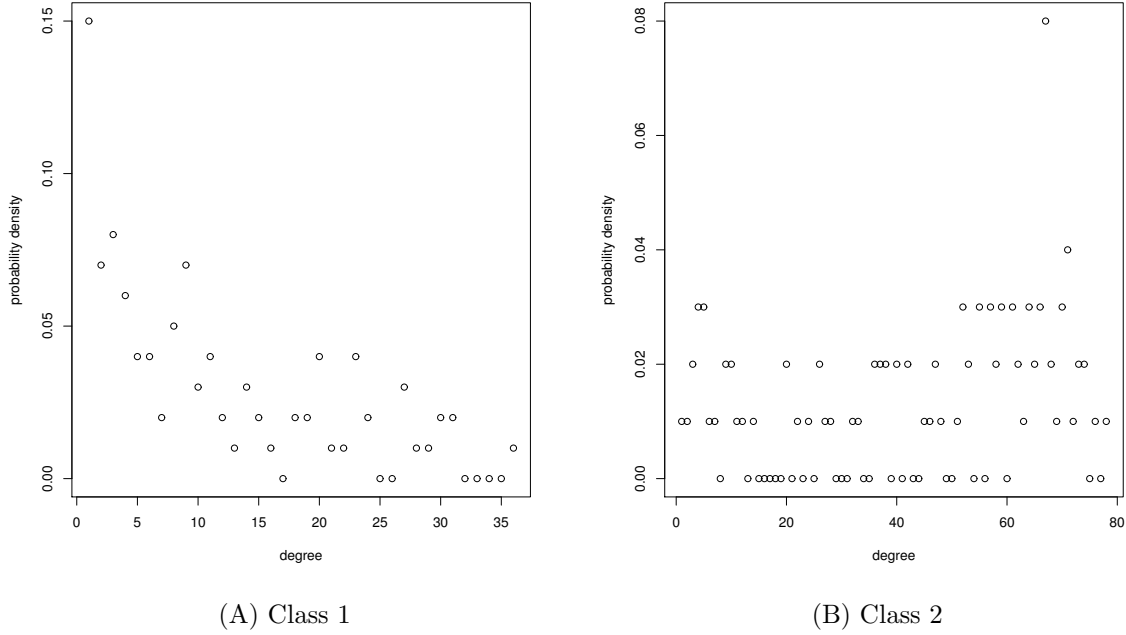
We perform P-DCov test with FDR control on all pairs of stocks and study the dependence between stocks conditional on the Fama-French three-factors. Under significance level $\alpha = 0.01$, we found out that 15.46% of the pairs of stocks are conditionally dependent given the three factors, which implies that the three factors may not be sufficient to explain the dependencies among stocks. As a comparison, we also implemented the conditional independence test with the distance covariance based test replaced by Pearson correlation based test. It turns out the 9.34% of the pairs are significant under the same significance level. This shows the P-DCov test is more powerful in discovering significant conditionally dependent pairs than the Pearson correlation test.

We then investigate the top 5 pairs of stocks that correspond to the largest test statistic values using the P-DCov test. They are (BHI, SLB), (CVX, XOM), (HAL, SLB), (COP, CVX), and (BHI, HAL). Interestingly, all six stocks involved are closely related to the oil industry. This reveals the high level of dependence among oil industry stocks that cannot be well explained by the Fama-French three-factor model. In addition, we examine the stock pairs that are conditionally dependent under the P-DCov test but not under the Pearson correlation test. The two most significant pairs are (C, USB) and (MRK, PFE). The first pair are in the financial industry (Citigroup and U.S. Bancorp) and the second pair are pharmaceutical companies (Merck & Co. and Pfizer). This shows that by using the proposed P-DCov, some interesting conditional dependence structures could be recovered. This is consistent with the findings that the sector correlations are still present even after adjusting Fama-French factors and 10 industrial factors (Fan *et al.*, 2016a).

5.2. Breast Cancer Data

In this section, we explore the difference in genetic networks between breast cancer patients who achieve pathologic Complete Response (pCR) and patients who do not achieve pCR. Achieving pCR is defined as no invasive and no in situ residuals left in breast in the surgical specimen. As studied in Kuerer *et al.* (1999) and Kaufmann *et al.* (2006), pCR has predicted long-term outcome in several neoadjuvant studies and hence serves as a potential surrogate marker for survival. In this study, we use the normalized gene expression data of 130 patients with stages I-III breast cancers analyzed by Hess *et al.* (2006). Among the 130 patients, 33 of them achieved pCR (class 1), while the other 97 patients did not achieve pCR (class 2). To construct the conditional dependence network for each class, we first perform a two-sample t -test between the two groups and select the 100 genes with the smallest p -values. Afterwards, we construct networks of these 100 selected genes for each class using P-DCov with the FDR control at level $\alpha = 0.01$. Notice, in this case, $d = 100$ and the corresponding sample sizes in two groups are $n_1 = 33$ and $n_2 = 97$ respectively.

In networks, the degree of a particular node describes how many edges are connected to this node, and the average degree serves as a measure of connectivity of the graphs. In Figure 6, we summarize the distribution of degrees for the genetic networks of class 1 and class 2 respectively. We see that the average degree of genetic network for class 1 is 9.7 which is much smaller than the average degree of

Fig. 6. Degree distribution of the genetic networks

network for class 2, which is 44.88. To look at the networks more closely, we select 7 genes among the 100 genes and draw the corresponding networks in Figure 7. We see that for Class 1 where pCR is achieved, gene MCCC2 is a hub and is connected with three other genes. However, in the network for Class 2, gene MCCC2 is disconnected from the other six genes. On the other hand, gene MAPT is isolated in the network for Class 1, but is connected with two other genes in class 2.

These findings imply that the two classes may have very different conditional dependence structures, and hence likely to have different precision matrices. As a result, when classification is the target, linear discriminant analysis may be too simple to capture the actual decision boundary.

6. Discussion

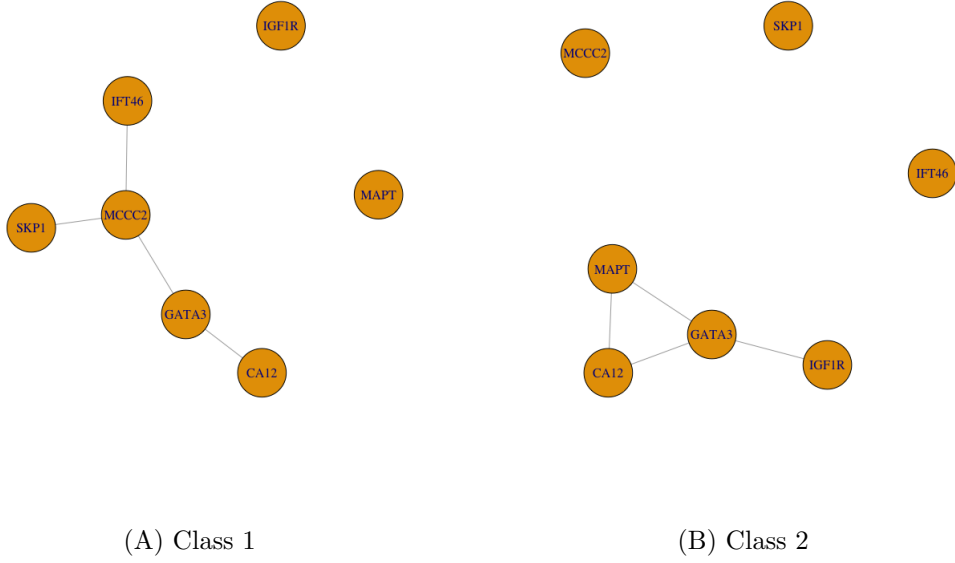
In this work, we proposed a general framework for testing conditional independence via projection and showed a new way to create dependency graphs. The current theoretical results assume that contribution of factors is sparse linear. How to extend the theory to the case of sparse additive model projection would be an interesting future work. Another interesting direction is to use the proposed test to create dependency graphs among groups of nodes, which could have applications in genetics.

An R package `pgraph` for implementing the proposed methodology is available on CRAN.

A. Proofs

LEMMA 3. *Under Condition 3, we have $\max_{i,j} \|(\mathbf{B}_x - \widehat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j)\| = O_p(a_n)$ and $\mathbb{E} \max_{i,j} \|(\mathbf{B}_x - \widehat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j)\| = O(a_n)$.*

PROOF. From Condition 3, it is obvious that $\max_{i,j} \|(\mathbf{B}_x - \widehat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j)\| = O_p(a_n)$. Let $U_n =$

Fig. 7. Genetic networks for the two classes based on 7 selected genes

$\max_{i,j} \|(\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j)\|$ and $\tilde{U}_n = U_n/a_n$. Then, we have

$$\begin{aligned} \mathbb{E}(\tilde{U}_n) &= \int_0^\infty \mathbb{P}(\tilde{U}_n > t) dt \\ &= \int_0^1 \mathbb{P}(\tilde{U}_n > t) dt + \int_1^\infty \mathbb{P}(\tilde{U}_n > t) dt \\ &\leq 1 + \int_1^\infty C_1^{-t} dt < \infty. \end{aligned}$$

As a result, the lemma is proved.

For the remaining proofs, we apply Taylor expansion to $\|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\|$ at $\epsilon_{i,x} - \epsilon_{j,x}$ and get

$$\begin{aligned} \|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\| &= \|\epsilon_{i,x} - \epsilon_{j,x}\| + \frac{\mathbf{c}_{i,j,x}^\top}{\|\mathbf{c}_{i,j,x}\|} (\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j) = \|\epsilon_{i,x} - \epsilon_{j,x}\| + D_{i,j,x}, \\ \|\hat{\epsilon}_{i,y} - \hat{\epsilon}_{j,y}\| &= \|\epsilon_{i,y} - \epsilon_{j,y}\| + \frac{\mathbf{c}_{i,j,y}^\top}{\|\mathbf{c}_{i,j,y}\|} (\mathbf{B}_y - \hat{\mathbf{B}}_y)(\mathbf{f}_i - \mathbf{f}_j) = \|\epsilon_{i,y} - \epsilon_{j,y}\| + D_{i,j,y}, \end{aligned} \quad (12)$$

where $\mathbf{c}_{i,j,x} = \lambda_{i,j,x}(\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}) + (1 - \lambda_{i,j,x})(\epsilon_{i,x} - \epsilon_{j,x})$ and $\mathbf{c}_{i,j,y} = \lambda_{i,j,y}(\hat{\epsilon}_{i,y} - \hat{\epsilon}_{j,y}) + (1 - \lambda_{i,j,y})(\epsilon_{i,y} - \epsilon_{j,y})$, for $\lambda_{i,j,x} \in [0, 1]$ and $\lambda_{i,j,y} \in [0, 1]$.

PROOF OF THEOREM 1. Using the Taylor expansion in (12), we have the following decomposition

$$\mathcal{V}_n^2(\hat{\epsilon}_x, \hat{\epsilon}_y) - \mathcal{V}_n^2(\epsilon_x, \epsilon_y) = T_1 + T_2 + T_3,$$

where

$$T_1 = \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,x} \|\epsilon_{i,y} - \epsilon_{j,y}\| + \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,x} \frac{1}{n^2} \sum_{k,l=1}^n \|\epsilon_{k,y} - \epsilon_{l,y}\| - \frac{2}{n^3} \sum_{i=1}^n \sum_{j,k=1}^n D_{i,j,x} \|\epsilon_{i,y} - \epsilon_{k,y}\|, \quad (13)$$

$$T_2 = \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,y} \|\epsilon_{i,x} - \epsilon_{j,x}\| + \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,y} \frac{1}{n^2} \sum_{k,l=1}^n \|\epsilon_{k,x} - \epsilon_{l,x}\| - \frac{2}{n^3} \sum_{i=1}^n \sum_{j,k=1}^n D_{i,j,y} \|\epsilon_{i,x} - \epsilon_{k,x}\|, \quad (14)$$

$$T_3 = \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,x} D_{i,j,y} + \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,x} \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,y} - \frac{2}{n^3} \sum_{i=1}^n \sum_{j,k=1}^n D_{i,j,x} D_{i,k,y}. \quad (15)$$

By Condition 3, we have $\max_{i,j} |D_{i,j,x}| = O_p(a_n \log n)$. Therefore,

$$\begin{aligned} |T_1| &\leq O_p(a_n \log n) \left(\frac{4}{n^2} \sum_{i,j=1}^n \|\epsilon_{i,y} - \epsilon_{j,y}\| \right), \\ |T_2| &\leq O_p(a_n \log n) \left(\frac{4}{n^2} \sum_{i,j=1}^n \|\epsilon_{i,x} - \epsilon_{j,x}\| \right), \\ |T_3| &\leq O_p((a_n \log n)^2). \end{aligned}$$

Another fact we easily observe is that: $n^{-2} \sum_{i,j=1}^n \|\epsilon_{i,x} - \epsilon_{j,x}\| = O_p(1)$, since $\mathbb{E}\|\epsilon_{i,x} - \epsilon_{j,x}\|$ is uniformly bounded over all (i, j) pairs and so is $\mathbb{E}(n^{-2} \sum_{i,j=1}^n \|\epsilon_{i,x} - \epsilon_{j,x}\|)$.

As a result, we know $\mathcal{V}_n^2(\hat{\epsilon}_x, \hat{\epsilon}_y) - \mathcal{V}_n^2(\epsilon_x, \epsilon_y) \xrightarrow{P} 0$. This combined with Lemma 1 leads to

$$\mathcal{V}_n^2(\hat{\epsilon}_x, \hat{\epsilon}_y) \xrightarrow{P} \mathcal{V}^2(\epsilon_x, \epsilon_y).$$

Remark: The result of Theorem 1 cannot be implied from that of Theorem 2, since independence between ϵ_x and ϵ_y is not assumed.

LEMMA 4. *For the $\mathbf{c}_{i,j,x}$ and $\mathbf{c}_{i,j,y}$ defined in (12), we have the following approximation error bound on the normalized version.*

$$\left\| \frac{\mathbf{c}_{i,j,x}}{\|\mathbf{c}_{i,j,x}\|} - \frac{\epsilon_{i,x} - \epsilon_{j,x}}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \right\| \leq \frac{2}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \max_{i,j} \|(\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j)\|, \quad (16)$$

$$\left\| \frac{\mathbf{c}_{i,j,y}}{\|\mathbf{c}_{i,j,y}\|} - \frac{\epsilon_{i,y} - \epsilon_{j,y}}{\|\epsilon_{i,y} - \epsilon_{j,y}\|} \right\| \leq \frac{2}{\|\epsilon_{i,y} - \epsilon_{j,y}\|} \max_{i,j} \|(\mathbf{B}_y - \hat{\mathbf{B}}_y)(\mathbf{f}_i - \mathbf{f}_j)\|. \quad (17)$$

PROOF. It suffices to show (16). First, we will show

$$\left\| \frac{\mathbf{c}_{i,j,x}}{\|\mathbf{c}_{i,j,x}\|} - \frac{\epsilon_{i,x} - \epsilon_{j,x}}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \right\| \leq \left\| \frac{\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}}{\|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\|} - \frac{\epsilon_{i,x} - \epsilon_{j,x}}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \right\|. \quad (18)$$

Denote by α_1 and α_2 the angle between $\mathbf{c}_{i,j,x}$ and $\epsilon_{i,x} - \epsilon_{j,x}$, and the angle between $\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}$ and $\epsilon_{i,x} - \epsilon_{j,x}$, respectively. It is easy to see that $0 \leq \alpha_1 \leq \alpha_2 \leq \pi$, and hence $\cos \alpha_1 \geq \cos \alpha_2$. By cosine formula,

$$\left\| \frac{\mathbf{c}_{i,j,x}}{\|\mathbf{c}_{i,j,x}\|} - \frac{\epsilon_{i,x} - \epsilon_{j,x}}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \right\|^2 = 2 - 2 \cos \alpha_1, \text{ and } \left\| \frac{\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}}{\|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\|} - \frac{\epsilon_{i,x} - \epsilon_{j,x}}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \right\|^2 = 2 - 2 \cos \alpha_2.$$

Therefore, (18) is proved and it remains to show that

$$\left\| \frac{\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}}{\|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\|} - \frac{\epsilon_{i,x} - \epsilon_{j,x}}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \right\| \leq \frac{2}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \max_{i,j \in \{1, \dots, n\}} \|(\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j)\|. \quad (19)$$

Left hand side of (19) can be rewritten as

$$\begin{aligned} &\left\| \frac{\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}}{\|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\|} - \frac{\epsilon_{i,x} - \epsilon_{j,x}}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \right\| \\ &= \left\| \frac{[(\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}) - (\epsilon_{i,x} - \epsilon_{j,x})] \|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\| - (\|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\| - \|\epsilon_{i,x} - \epsilon_{j,x}\|)(\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x})}{\|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\| \|\epsilon_{i,x} - \epsilon_{j,x}\|} \right\| \\ &\leq \frac{1}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} (\|(\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}) - (\epsilon_{i,x} - \epsilon_{j,x})\| + \|\hat{\epsilon}_{i,x} - \hat{\epsilon}_{j,x}\| - \|\epsilon_{i,x} - \epsilon_{j,x}\|) \\ &\leq \frac{2}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \max_{i,j \in \{1, \dots, n\}} \|(\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j)\|. \end{aligned}$$

Combining (18) and (19), the lemma is proved.

LEMMA 5. Under Conditions 1 and 2, and the null hypothesis that $\epsilon_x \perp\!\!\!\perp \epsilon_y$, for any $\gamma > 0$,

$$\frac{1}{n^\gamma \log n} \left[\frac{1}{n^2} \sum_{i,j=1}^n \frac{1}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \right] \xrightarrow{P} 0, \quad \frac{1}{n^\gamma \log n} \left[\frac{1}{n^2} \sum_{i,j=1}^n \frac{1}{\|\epsilon_{i,y} - \epsilon_{j,y}\|} \right] \xrightarrow{P} 0.$$

PROOF. We will only show the first result involving ϵ_x with the other one follows similarly. For any $\delta > 0$, let

$$R_n = \frac{1}{n^2} \sum_{i,j=1}^n \frac{1}{\|\epsilon_{i,x} - \epsilon_{j,x}\|}, \quad \bar{R}_n = \frac{1}{n^2} \sum_{i,j=1}^n \frac{1}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \wedge n^{2+\delta}.$$

Then for $\forall \epsilon > 0$,

$$\mathbb{P}[|R_n - \bar{R}_n| > \epsilon] \leq n^2 \mathbb{P}[\|\epsilon_{i,x} - \epsilon_{j,x}\| < n^{-2-\delta}] \leq C n^2 n^{-2-\delta} = C n^{-\delta}, \quad (20)$$

due to the Condition 2 that the density function of $\|\epsilon_{i,x} - \epsilon_{j,x}\|$ is pointwise bounded. Therefore, $|R_n - \bar{R}_n| \xrightarrow{P} 0$, which leads to

$$\left| \frac{R_n}{n^\gamma \log n} - \frac{\bar{R}_n}{n^\gamma \log n} \right| \xrightarrow{P} 0. \quad (21)$$

On the other hand,

$$\begin{aligned} \mathbb{E}\left[\frac{1}{\log n} \frac{1}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} \wedge n^{2+\delta}\right] &= \frac{1}{\log n} \mathbb{P}\left(\frac{1}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} > n^{2+\delta}\right) n^{2+\delta} + \frac{1}{\log n} \int_{n^{-2-\delta}}^{\infty} \frac{1}{t} h_{\|\epsilon_{i,x} - \epsilon_{j,x}\|}(t) dt \\ &\leq \frac{C}{\log n} + \frac{1}{\log n} \int_{n^{-2-\delta}}^{C_0} \frac{1}{x} h_x(x) dx + \frac{1}{\log n} \int_{C_0}^{\infty} \frac{1}{t} h_{\|\epsilon_{i,x} - \epsilon_{j,x}\|}(t) dt \\ &\leq \frac{C}{\log n} + \frac{C}{\log n} \int_{n^{-2-\delta}}^{C_0} \frac{1}{x} dx + \frac{1}{C_0 \log n} \mathbb{P}(\|\epsilon_{i,x} - \epsilon_{j,x}\| > C_0) \\ &\leq \frac{C}{\log n} + \frac{C}{\log n} [\log(C_0) + \log(n^{2+\delta})] + \frac{1}{C_0 \log n} \\ &\leq \frac{C}{\log n} + C' + \frac{1}{C_0 \log n}, \end{aligned} \quad (22)$$

where $h_{\|\epsilon_{i,x} - \epsilon_{j,x}\|}$ is the density of $\|\epsilon_{i,x} - \epsilon_{j,x}\|$. In the above derivation, the first inequality can be easily seen from (20) and the second inequality utilizes Condition 2.

Therefore, $\bar{R}_n / \log n$ is bounded in L_1 and since $n^\gamma \rightarrow \infty$, $\bar{R}_n / [n^\gamma \log(n)]$ converges to 0 in L_1 and hence in probability, i.e.,

$$\frac{\bar{R}_n}{n^\gamma \log(n)} \xrightarrow{P} 0. \quad (23)$$

This, combined with (21) yields

$$\frac{R_n}{n^\gamma \log(n)} \xrightarrow{P} 0. \quad (24)$$

This completes the proof of Lemma 5.

To prove Theorem 2, we first introduce two propositions.

PROPOSITION 1. Under Conditions 1 and 2, and the null hypothesis that $\epsilon_x \perp\!\!\!\perp \epsilon_y$,

$$T_1 = O_p(a_n/n), \quad T_2 = O_p(a_n/n)$$

PROOF. From (13), we rewrite T_1 as

$$\begin{aligned} T_1 &= \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,x} \left(\|\epsilon_{i,y} - \epsilon_{j,y}\| + \frac{1}{n^2} \sum_{k,l=1}^n \|\epsilon_{k,y} - \epsilon_{l,y}\| - \frac{1}{n} \sum_{k=1}^n \|\epsilon_{i,y} - \epsilon_{k,y}\| - \frac{1}{n} \sum_{k=1}^n \|\epsilon_{j,y} - \epsilon_{k,y}\| \right) \\ &\doteq \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,x} A_{i,j,y}, \end{aligned}$$

with $A_{i,j,y}$ self-defined by the equation.

Let us consider term

$$\mathbb{E}(T_1^2) = \frac{1}{n^4} \sum_{i \neq j, k \neq l} \mathbb{E}(D_{i,j,x} D_{k,l,x} A_{i,j,y} A_{k,l,y}) = \frac{1}{n^4} \sum_{i \neq j, k \neq l} \mathbb{E}(D_{i,j,x} D_{k,l,x}) \mathbb{E}(A_{i,j,y} A_{k,l,y}). \quad (25)$$

We can separate the above quantity into three parts. It is easy to see that $D_{i,j,x}$ are identically distributed with respect to different pairs of (i, j) when $i \neq j$. Let us define the following three sets of index quadruples:

- $I_1 = \{(i, j, k, l) | \text{there are four distinct values in } \{i, j, k, l\}\}.$
- $I_2 = \{(i, j, k, l) | i \neq j, k \neq l, \text{ and there are three distinct values in } \{i, j, k, l\}\}.$
- $I_3 = \{(i, j, k, l) | i \neq j, k \neq l, \text{ and there are two distinct values in } \{i, j, k, l\}\}.$

Let us suppose $\mathbb{E}(D_{i,j,x} D_{k,l,x}) = c_1$, for $(i, j, k, l) \in I_1$; $\mathbb{E}(D_{i,j,x} D_{k,l,x}) = c_2$, for $(i, j, k, l) \in I_2$. $\mathbb{E}(D_{i,j,x} D_{k,l,x}) = c_3$, for $(i, j, k, l) \in I_3$. By Condition 3, we know c_1, c_2 and c_3 are all of order $O(a_n^2)$. Also, $\mathbb{E}(A_{i,j,y}) = O(1)$. Then we have

$$\mathbb{E}(T_1^2) = \mathbb{E} \left(\frac{c_1}{n^4} \sum_{I_1} A_{i,j,y} A_{k,l,y} + \frac{c_2}{n^4} \sum_{I_2} A_{i,j,y} A_{k,l,y} + \frac{c_3}{n^4} \sum_{I_3} A_{i,j,y} A_{k,l,y} \right). \quad (26)$$

On the other hand, we observe that $\sum_{j=1}^n A_{i,j,y} = 0$ by definition and $A_{i,j,y} = A_{j,i,y}$, so we have

$$\sum_{I_2} A_{i,j,y} A_{k,l,y} = \sum_{i=1}^n \left(\sum_{j=1}^n A_{i,j,y} \right)^2 - \sum_{i=1}^n \sum_{j=1}^n A_{i,j,y}^2 = - \sum_{i=1}^n \sum_{j=1}^n A_{i,j,y}^2.$$

By Condition 1, we know all the second order terms of distances of differences ($\|\epsilon_{i,y} - \epsilon_{j,y}\|^2, \|\epsilon_{i,y} - \epsilon_{j,y}\| \cdot \|\epsilon_{i,y} - \epsilon_{k,y}\|$ as examples) have bounded expectation, and thus all the second order terms of $A_{i,j,y}$'s also have bounded expectations. Therefore, $\mathbb{E}(n^{-4} \sum_{I_3} A_{i,j,y} A_{k,l,y}) = O(n^{-2})$. Finally, since $\sum_{i=1}^n \sum_{j=1}^n A_{i,j,y} = 0$,

$$\begin{aligned} \sum_{I_1} A_{i,j,y} A_{k,l,y} &= \left(\sum_{i=1}^n \sum_{j=1}^n A_{i,j,y} \right)^2 - \sum_{I_2} A_{i,j,y} A_{k,l,y} - \sum_{I_3} A_{i,j,y} A_{k,l,y} - \sum_{i=1}^n A_{i,i,y}^2 \\ &= - \sum_{I_2} A_{i,j,y} A_{k,l,y} - \sum_{I_3} A_{i,j,y} A_{k,l,y} - \sum_{i=1}^n A_{i,i,y}^2. \end{aligned}$$

This combined with our previous calculations leads to $\mathbb{E}(n^{-4} \sum_{I_1} A_{i,j,y} A_{k,l,y}) = O(n^{-2})$. As a result, we have $\mathbb{E}(T_1^2) = O(a_n^2/n^2)$. Together with Chebychev's inequality, we know $T_1^2 = O_p(a_n^2/n^2)$ and equivalently, $T_1 = O_p(a_n/n)$. Similarly, we could show that $T_2 = O_p(a_n/n)$.

PROPOSITION 2. Under Conditions 1, 2, 3, and 4, and the null hypothesis that $\epsilon_x \perp \epsilon_y$,

$$T_3 = O_p\left(\frac{a_n^2 \log K}{\sqrt{n}}\right).$$

PROOF. Recall that

$$\begin{aligned} T_3 &= \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,x} D_{i,j,y} + \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,x} \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,y} - \frac{2}{n^3} \sum_{i=1}^n \sum_{j,k=1}^n D_{i,j,x} D_{i,k,y} \\ &\doteq \frac{1}{n^2} \sum_{i,j=1}^n D_{i,j,x} B_{i,j,y}, \end{aligned}$$

with $B_{i,j,y}$ self-defined in the above equation. We can easily see that $\sum_{i=1}^n B_{i,j,y} = 0$, for any j . Let $B_{\max} = \max_{i,j} |B_{i,j,y}|$, then we define $\tilde{B}_{i,j,y} = B_{i,j,y}/(2B_{\max}) + 0.5$. In this way, we know $\tilde{B}_{i,j,y} \in [0, 1]$ and $\sum_{i=1}^n \tilde{B}_{i,j,y} = 1/2$ for any j . By Condition 3, we know that $B_{\max} = O_p(a_n)$.

Then we can rewrite T_3 in the following form:

$$T_3 = \frac{2B_{\max}}{n^2} \sum_{i,j=1}^n D_{i,j,x} \tilde{B}_{i,j,y} - \frac{B_{\max}}{n^2} \sum_{i,j=1}^n D_{i,j,x} \doteq T_{31} - T_{32}.$$

Let us look at T_{31} first. If we denote \mathbf{D} and $\tilde{\mathbf{B}}$ as the matrix of dimension $n \times n$ composed of elements $D_{i,j,x}$ and $\tilde{B}_{i,j,y}$, we know that

$$|T_{31}| \leq \frac{2B_{\max}}{n^2} \|\mathbf{D}\|_F \|\tilde{\mathbf{B}}\|_F = O_p(a_n/n^2) O_p(a_n n) O_p(\sqrt{n}) = O_p\left(\frac{a_n^2}{\sqrt{n}}\right). \quad (27)$$

Then, let us proceed to term T_{32} . Here, we write $D_{i,j,x}$ in another form as a sum of two terms and bound them separately.

$$\begin{aligned} D_{i,j,x} &= \frac{(\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x})^\top}{\|\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x}\|} (\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j) + \left(\frac{\mathbf{c}_{i,j,x}}{\|\mathbf{c}_{i,j,x}\|} - \frac{\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x}}{\|\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x}\|} \right) (\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j) \\ &\equiv \frac{(\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x})^\top}{\|\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x}\|} (\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j) + r_{i,j,x}. \end{aligned} \quad (28)$$

As a result, we know

$$T_{32} = \frac{B_{\max}}{n^2} \sum_{i,j=1}^n r_{i,j,x} + \frac{B_{\max}}{n^2} \sum_{i,j=1}^n \frac{(\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x})^\top}{\|\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x}\|} (\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j).$$

By Lemma 4, we know

$$|r_{i,j,x}| \leq \max_{i,j} \|(\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j)\|^2 \frac{2}{\|\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x}\|}, \quad (29)$$

where $\max_{i,j} \|(\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j)\|^2 = O_p(a_n^2)$.

So the first term in T_{32} has rate $n^{-2} B_{\max} \sum_{i,j=1}^n r_{i,j,x} = O_p(a_n^3 (\log n) n^\gamma)$.

The second term in T_{32} can be rewritten in terms of trace:

$$\begin{aligned} \left\| \frac{B_{\max}}{n^2} \sum_{i,j=1}^n \frac{(\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x})^\top}{\|\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x}\|} (\mathbf{B}_x - \hat{\mathbf{B}}_x)(\mathbf{f}_i - \mathbf{f}_j) \right\| &= \left| B_{\max} \text{Tr} \left((\mathbf{B}_x - \hat{\mathbf{B}}_x) \frac{1}{n^2} \sum_{i,j=1}^n (\mathbf{f}_i - \mathbf{f}_j) \frac{(\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x})^\top}{\|\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x}\|} \right) \right| \\ &\doteq \left| B_{\max} \text{Tr} \left((\mathbf{B}_x - \hat{\mathbf{B}}_x) \mathbf{W} \right) \right|, \\ &\leq B_{\max} \sum_{l=1}^p \|\mathbf{B}_{x,l} - \hat{\mathbf{B}}_{x,l}\|_1 \max_{i,j} |W(i,j)|, \end{aligned} \quad (30)$$

where \mathbf{W} is self-defined and $W(i,j)$ is the element on the i -th row and j -column of matrix \mathbf{W} . Let us take $(i,j) = (1,1)$ as an example, and look at $W(1,1) = \frac{1}{n^2} \sum_{i,j=1}^n (f_{i,1} - f_{j,1}) \frac{\boldsymbol{\epsilon}_{i,x,1} - \boldsymbol{\epsilon}_{j,x,1}}{\|\boldsymbol{\epsilon}_{i,x} - \boldsymbol{\epsilon}_{j,x}\|}$. We easily see

that $\mathbb{E}W(1, 1) = 0$, due to facts: $\epsilon_{i,x}$ and $\epsilon_{j,x}$ are mutually independent of \mathbf{f} with any observation indices; and $\mathbb{E}[(\epsilon_{i,x} - \epsilon_{j,x})/\|\epsilon_{i,x} - \epsilon_{j,x}\|] = 0$. Furthermore,

$$\mathbb{E}(W(1, 1)^2) = \frac{1}{n^4} \sum_{i,j,k,l=1}^n (f_{i,1} - f_{j,1}) \frac{\epsilon_{i,x,1} - \epsilon_{j,x,1}}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} (f_{k,1} - f_{l,1}) \frac{\epsilon_{k,x,1} - \epsilon_{l,x,1}}{\|\epsilon_{k,x} - \epsilon_{l,x}\|}.$$

Similar to the reasoning in Proposition 1, we have n^4 terms in I_1 . But in this scenario, $\mathbb{E}(f_{i,1} - f_{j,1}) \frac{\epsilon_{i,x,1} - \epsilon_{j,x,1}}{\|\epsilon_{i,x} - \epsilon_{j,x}\|} (f_{k,1} - f_{l,1}) \frac{\epsilon_{k,x,1} - \epsilon_{l,x,1}}{\|\epsilon_{k,x} - \epsilon_{l,x}\|} = 0$ due to independence, therefore we know

$$\mathbb{E}(W(1, 1)^2) = O(1/n).$$

As a result, we know $|W(1, 1)| = O_p(n^{-1/2})$, and thus $\max_{i,j} |W(i, j)| = O_p(n^{-1/2} \log K)$. Furthermore, we can bound the term in (30) with rate $O_p(n^{-1/2} a_n e_n \log K)$.

Combining T_{31} and T_{32} , we know $T_3 = O_p\{a_n^3(\log n)n^\gamma\} \vee (n^{-1/2} a_n e_n \log K)$ and the proposition is proved by looking at Conditions 3 and 4.

PROOF OF THEOREM 2. Recall the notations we used in the proof of Theorem 1,

$$\mathcal{V}_n^2(\hat{\epsilon}_x, \hat{\epsilon}_y) - \mathcal{V}_n^2(\epsilon_x, \epsilon_y) = T_1 + T_2 + T_3.$$

By Propositions 1 and 2, we have for any $\gamma > 0$,

$$n(T_1 + T_2 + T_3) = O_p(a_n) + O_p\{(n^{1+\gamma}(\log n)a_n^3) \vee (a_n e_n \log K \sqrt{n})\}.$$

Combined with Lemma 2, the theorem is proved.

PROOF OF COROLLARY 2. The result follows directly from the proofs of Theorems 1 and 2 and an application of Slutsky's theorem.

PROOF OF THEOREM 3. The proof of Theorem 3 follows similarly as Theorem 6 in Székely *et al.* (2007). Here we omit the details for brevity.

References

- Anderson, T. W. (1962) *An introduction to multivariate statistical analysis*. Wiley New York.
- Belloni, A., Chernozhukov, V. *et al.* (2011) L1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, **39**, 82–130.
- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.
- Blomqvist, N. (1950) On a measure of dependence between two random variables. *The Annals of Mathematical Statistics*, 593–600.
- Blum, J., Kiefer, J. and Rosenblatt, M. (1961) Distribution free tests of independence based on the sample distribution function. *The Annals of Mathematical Statistics*, 485–498.
- Bühlmann, P. and Van De Geer, S. (2011) *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.
- Cai, T., Liu, W. and Luo, X. (2011) A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, **106**, 594–607.

- Cai, T. T., Li, H., Liu, W. and Xie, J. (2013) Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, **100**, 139–156.
- Dempster, A. P. (1972) Covariance selection. *Biometrics*, 157–175.
- Drton, M. and Perlman, M. D. (2004) Model selection for gaussian concentration graphs. *Biometrika*, **91**, 591–602.
- Edwards, D. (2000) *Introduction to graphical modelling*. Springer.
- Fama, E. F. and French, K. R. (1993) Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, **33**, 3–56.
- Fan, J., Feng, Y. and Song, R. (2011) Nonparametric independence screening in sparse ultra-high dimensional additive models. *Journal of the American Statistical Association*, **106**, 544–557.
- Fan, J., Feng, Y. and Wu, Y. (2009) Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, **3**, 521.
- Fan, J., Furger, A. and Xiu, D. (2016a) Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high frequency data. *Journal of Business & Economic Statistics*, 489–503.
- Fan, J., Li, Q. and Wang, Y. (2016b) Robust estimation of high-dimensional mean regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, to appear.
- Finegold, M. A. and Drton, M. (2009) Robust graphical modeling with t-distributions. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 169–176. AUAI Press.
- Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.
- Hastie, T., Tibshirani, R. and Wainwright, M. (2015) *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.
- Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J. *et al.* (2006) Pharmacogenomic predictor of sensitivity to pre-operative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, **24**, 4236–4244.
- Hoeffding, W. (1948) A non-parametric test of independence. *The Annals of Mathematical Statistics*, 546–557.
- Hollander, M., Wolfe, D. A. and Chicken, E. (2013) *Nonparametric statistical methods*. John Wiley & Sons.
- Kaufmann, M., Hortobagyi, G. N., Goldhirsch, A., Scholl, S., Makris, A., Valagussa, P., Blohmer, J.-U., Eiermann, W., Jackesz, R., Jonat, W. *et al.* (2006) Recommendations from an international expert panel on the use of neoadjuvant (primary) systemic treatment of operable breast cancer: an update. *Journal of Clinical Oncology*, **24**, 1940–1949.

- Kuerer, H. M., Newman, L. A., Smith, T. L., Ames, F. C., Hunt, K. K., Dhingra, K., Theriault, R. L., Singh, G., Binkley, S. M., Sneige, N. *et al.* (1999) Clinical course of breast cancer patients with complete pathologic primary tumor and axillary lymph node response to doxorubicin-based neoadjuvant chemotherapy. *Journal of Clinical Oncology*, **17**, 460–460.
- Lauritzen, S. L. (1996) *Graphical models*. Oxford University Press.
- Linton, O. and Gozalo, P. (1997) Conditional independence restrictions: testing and estimation. *V Cowles Foundation Discussion Paper*, **1140**.
- Liu, H., Lafferty, J. and Wasserman, L. (2009) The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, **10**, 2295–2328.
- Meinshausen, N. and Bühlmann, P. (2006) High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 1436–1462.
- Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009) Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 1009–1030.
- Stock, J. H. and Watson, M. W. (2002) Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, **20**, 147–162.
- Stone, C. J. (1985) Additive regression and other nonparametric models. *The Annals of Statistics*, **13**, 689–705.
- Su, L. and White, H. (2007) A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, **141**, 807–834.
- Su, L. and White, H. (2008) A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, **24**, 829–864.
- Su, L. and White, H. (2014) Testing conditional independence via empirical likelihood. *Journal of Econometrics*.
- Székely, G. J., Rizzo, M. L., Bakirov, N. K. *et al.* (2007) Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, **35**, 2769–2794.
- Wainwright, M. J. and Jordan, M. I. (2008) Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, **1**, 1–305.
- Wang, L. (2013) The l1 penalized lad estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, **120**, 135–151.
- Wilks, S. (1935) On the independence of k sets of normally distributed statistical variables. *Econometrica, Journal of the Econometric Society*, 309–326.
- Xue, L. and Zou, H. (2012) Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, **40**, 2541–2571.